Adaptive Long-Read Sequencing Reveals GGC Repeat Expansion in *ZFHX3* Associated with Spinocerebellar Ataxia Type 4

Zhongbo Chen, BM BCh, MA, PhD,[1,2,3†] Emil K. Gustavsson, PhD,[2,3†] Hannah Macpherson, MSc,[1,2]
Claire Anderson, PhD,[2,3] Chris Clarkson, PhD,[4] Clarissa Rocca, MSc,[5] Eleanor Self, BSc,[5] Pilar Alvarez Jerez, MSc,[1,6]
Annarita Scardamaglia, BSc,[5] David Pellerin, MD,[5] Kylie Montgomery, BSc,[2,3] Jasmaine Lee, BSc,[5]
Delia Gagliardi, MD,[5] Huihui Luo, MSc,[5] Genomics England Research Consortium, John Hardy, PhD, FMedSci,[1,7,8,9,10]
James Polke, PhD,[11] Andrew B. Singleton, PhD,[6,12] Cornelis Blauwendraat, PhD,[6,12] Katherine D. Mathews, MD,[13,14]
Arianna Tucci, MD, PhD,[4] Ying-Hui Fu, PhD,[15,16,17,18†] Henry Houlden, MD, PhD,[5,11*] Mina Ryten, MD, PhD,[2,3†] and
Louis J. Ptáček, MD, PhD[15,16,17,18†]

[1]Department of Neurodegenerative Disease, Queen Square Institute of Neurology, University College London, London, United Kingdom
[2]Department of Genetics and Genomic Medicine, Great Ormond Street Institute of Child Health, University College London, London, United Kingdom
[3]NIHR Great Ormond Street Hospital Biomedical Research Centre, University College London, London, United Kingdom
[4]William Harvey Research Institute, Queen Mary University of London, London, United Kingdom
[5]Department of Neuromuscular Disease, Queen Square Institute of Neurology, University College London, London, United Kingdom
[6]Center for Alzheimer's and Related Dementias, National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA
[7]Reta Lila Weston Institute, Queen Square Institute of Neurology, University College London, London, United Kingdom
[8]UK Dementia Research Institute, University College London, London, United Kingdom
[9]NIHR University College London Hospitals Biomedical Research Centre, London, United Kingdom
[10]Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong, China
[11]The Neurogenetics Laboratory, National Hospital for Neurology and Neurosurgery, London, United Kingdom
[12]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, USA
[13]Department of Pediatrics, University of Iowa Carver College of Medicine, Iowa City, Iowa, USA
[14]Department of Neurology, University of Iowa Carver College of Medicine, Iowa City, Iowa, USA
[15]Department of Neurology, University of California San Francisco, San Francisco, California, USA
[16]Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA
[17]Weill Institute for Neuroscience, University of California San Francisco, San Francisco, California, USA
[18]Kavli Institute for Fundamental Neuroscience, University of California San Francisco, San Francisco, California, USA

ABSTRACT: Background: Spinocerebellar ataxia type 4 (SCA4) is an autosomal dominant ataxia with invariable sensory neuropathy originally described in a family with Swedish ancestry residing in Utah more than 25 years ago. Despite tight linkage to the 16q22 region, the molecular diagnosis has since remained elusive.

Objectives: Inspired by pathogenic structural variation implicated in other 16q-ataxias with linkage to the same locus, we revisited the index SCA4 cases from the Utah family using novel technologies to investigate structural variation within the candidate region.

Methods: We adopted a targeted long-read sequencing approach with adaptive sampling on the Oxford

Correspondence to: Prof. Henry Houlden, Department of Neuromuscular Disease, Queen Square Institute of Neurology, University College London, London WC1N 3BG, UK; E-mail: h.houlden@ucl.ac.uk

†These authors contributed equally to this study.

Nanopore Technologies (ONT) platform that enables the detection of segregating structural variants within a genomic region without *a priori* assumptions about any variant features.

Results: Using this approach, we found a heterozygous $(GGC)_n$ repeat expansion in the last coding exon of the zinc finger homeobox 3 (*ZFHX3*) gene that segregates with disease, ranging between 48 and 57 GGC repeats in affected probands. This finding was replicated in a separate family with SCA4. Furthermore, the estimation of this GGC repeat size in short-read whole genome sequencing (WGS) data of 21,836 individuals recruited to the 100,000 Genomes Project in the UK and our in-house dataset of 11,258 exomes did not reveal any pathogenic repeats, indicating that the variant is ultrarare.

Conclusions: These findings support the utility of adaptive long-read sequencing as a powerful tool to decipher causative structural variation in unsolved cases of inherited neurological disease. © 2024 The Authors. *Movement Disorders* published by Wiley Periodicals LLC on behalf of International Parkinson and Movement Disorder Society.

Key Words: ataxia; long-read sequencing; repeat expansion disorder

Hereditary ataxias are a clinically and genetically heterogeneous group of disorders characterized by progressive incoordination secondary to dysfunction of the cerebellum and its connections.[1-3] Advances in next-generation sequencing techniques have improved the diagnostic rate such that variants in more than 300 genes have now been causally implicated in ataxia.[1,3] Furthermore, recent developments in bioinformatics analyses and long-read sequencing have yielded novel characterization of late-onset ataxia cases associated with repeat expansions in *RFC1* and *FGF14*, causing cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS), and spinocerebellar ataxia (SCA) 27B, respectively.[4,5] Nonetheless, some ataxia cases have remained intriguingly elusive to precise molecular characterization despite tight linkage and deep phenotyping, features normally conducive to the identification of the genetic cause.

One such disorder is SCA4 (MIM 600223), an autosomal dominant ataxia associated with prominent axonal sensory neuropathy, which is almost uniformly the earliest sign on examination, with normal ocular movements. This syndrome was originally described in a large five-generation kindred of Swedish ancestry in Utah.[6] Within this pedigree, linkage to chromosome (chr) 16q22.1 was reported with a logarithm of the odds (LOD) score of 5.93 at the microsatellite marker DS16S397 (hg38: chr16:66704306–66,704,561).[6] Subsequently, a five-generation family in northern Germany was found to have a similar clinical presentation, with the candidate region narrowed to span an approximate 3.69 cM interval between D16S3019 and D16S512 (hg38: chr16:66095251–74033670).[7] Despite refining the causative region through linkage analysis, the pathogenic genetic variant remained unknown after screening for genes with neuronal expression and both CAG and non-CAG repeats within the candidate locus.[7,8] Of interest, several Japanese families presenting with pure ataxia in the absence of peripheral sensory impairment were also found to have linkage to 16q22, but further analyses revealed a pathogenic pentanucleotide TGGAA repeat expansion in *BEAN1* within that locus (now coined SCA31).[9] Recently, Chinese families presenting with ataxia without sensory involvement but with linkage to 16q22.1 were found to have a causative CAG repeat expansion in *THAP11*, highlighting another important example of 16q-ataxia, at the locus for SCA4.[10]

Inspired by the discovery of causative structural variation in other 16q-ataxias, we revisited the index cases in the original Utah family, hypothesizing that the elusive molecular cause of SCA4 is likely secondary to a complex structural variant. To investigate this, we adopted a targeted long-read sequencing approach with adaptive sampling on the Oxford Nanopore Technologies (ONT) platform, harnessing existing knowledge of the linkage within the 16q22 region. Using this approach, we found a heterozygous $(GGC)_n$ repeat expansion in the last coding exon of the zinc finger homeobox 3 (*ZFHX3*) gene that segregates with disease, ranging between 48 and 57 GGC repeats in affected probands. This finding was replicated in another patient with Swedish ancestry presenting with a SCA4-phenotype in a separate pedigree. Further estimation of this GGC repeat size in short-read whole genome sequencing (WGS) data of 21,836 individuals recruited to the 100,000 Genomes Project in the UK did not reveal any repeat expansions within the pathogenic range indicating that the GGC repeat expansion is a rare variant. Taken together, these findings support pathogenic GGC repeat expansion in *ZFHX3* as the underlying genetic cause of SCA4, within the index pedigree. Our study also supports the utility of long-read sequencing with adaptive sampling and provides a workflow to guide the investigation of such unsolved genetic disorders.

Patients and Methods

Patients and Participants

This study was approved by the Institutional Review Board for Human Research at the University of Utah

School of Medicine with written informed consent obtained from each participant. Venous blood had been taken previously and transformed into lymphoblastoid cell lines and frozen as part of the original study.[6] DNA was also extracted from whole blood lysate with phenol/chloroform extraction followed by isopropranolol precipitation in the original study.[6]

Lymphoblast Cell Culture and DNA Extraction

Cell lines of participants from the index SCA4 family were cultured in RPMI 1640 medium (Thermo Fisher), supplemented with 10% fetal bovine serum (Thermo Fisher). The cells were incubated in a 5% CO_2 air humidified cell culture incubator at 37 °C. The cell count was determined using automated cell counters (Thermo Fisher). Once the cell concentration reached 1×10^6 cells/ml, they were harvested by centrifugation at $100 \times g$ for 10 minutes at room temperature to retrieve the cell pellet. Genomic DNA was extracted from 1×10^6 cultured cells using the New England Biolabs Monarch High Molecular Weight DNA Extraction Kit for Cells and Blood standard protocol. Quality of both the newly extracted DNA from lymphoblast cell lines and DNA from the original study was checked using the Genomic DNA 165 kb kit for the Agilent Femto Pulse, NanoDrop, and Qubit. DNA from the original study ranged in size from 8 to 17 kb. High-molecular-weight DNA from lymphoblasts were sheared to 12–17 kb using the Diagenode Megaruptor three at a concentration of 15 ng/µL at speed setting 30. We pre-sheared samples that were too viscous using the Diagenode DNA fluid kit at speed setting 40 before running the usual protocol. Sheared samples were then assessed for fragment length on the Femto Pulse and concentrated through a 1X Promega ProNex bead cleanup. We selected seven samples overall that fulfilled criteria for quality and quantity of DNA required for long-read sequencing.

Targeted Long-Read DNA Sequencing Using Adaptive Sampling

Given the clear region of linkage, we used a targeted long-read sequencing approach through adaptive sampling (full protocol: https://www.protocols.io/view/native-barcoding-sqk-nbd114-gdna-for-adaptive-samp-kxygx3qx4g8j/v1). This allows the selection of DNA based on a target sequence, controlled computationally in real time.[11] For the library preparation, a total of 1200–1500 ng of each sample was inputted into the standard SQK-NBD114.24 ONT (Oxford, UK) protocol. An equimolar pool was made of the barcoded samples, and 44–56 fmol of this pool was loaded onto the flow cell. Sequencing of the seven samples was performed on the PromethION of four to five samples per flow cell (R10.4.1). The library was sequenced using

adaptive sampling through a custom browser extensible data (BED) file covering the region of interest. For the design of the BED file, we selected a genomic region of 20 Mb, which encompasses the region of linkage 16q22.1, extending downstream to 16q21 and 16q13, and upstream to 16q22.2, 16q22.3, and 16q23.1. This region was selected to ensure the inclusion of all previously described microsatellite markers of interest located between chr16:56,000,000–76,100,000 (hg38),[6,8] such that the region could be enriched in the sequencing.

Structural Variant and Short Tandem Repeat Expansion Detection

As we hypothesized that the molecular etiology of SCA4 is secondary to structural variation, we employed a structural variant calling protocol. High accuracy base calling was performed with Guppy (v.7.0.9, ONT). Resulting FASTQ files, with high accuracy reads (quality score >9), were processed through our Snakemake pipeline.[12] This pipeline generated fastq stats using NanoStat[13] (v.1.6), mapped the reads to the genome using minimap2[14] (v 2.26; RRID: SCR_018550), and used Sniffles[15] (v.2.0.7; RRID: SCR_017619) for calling structural variants. Our pipeline is available at https://github.com/egustavsson/long-read_SV_calling.git. A custom script was designed to identify candidate structural variants that segregate to disease. Candidate structural variants were subsequently viewed on Integrative Genome Viewer (IGV).[11]

Estimation of Repeat Size in 100,000 Genomes Project

We investigated whether the candidate structural variation of interest in SCA4 segregated between individuals presenting with ataxia and controls recruited in the 100,000 Genomes Project[16] and its distribution within unaffected individuals across different genetically determined ancestries. The 100,000 Genomes Project is a UK program to assess the value of WGS in patients with rare diseases, and participants were recruited with consent after approval from the national research ethics committee.[16] We used ExpansionHunter v.3.1.2[17] to estimate repeat sizes using short-read WGS data of 803 individuals presenting with ataxia, 14,186 individuals presenting with a neurological phenotype not associated with ataxia, and 7650 individuals presenting without a neurological complaint, recruited in the 100,000 Genomes Project. We applied the same genotyping analysis to 11,258 exomes from our internal database Koios, comprising patients with neurological and neurodevelopmental presentation recruited from around the world.

Functional Genomic Analysis of *ZFHX3* and 16q22

We leveraged the Human Protein Atlas (HPA; v.23.0)[18] to examine the expression of the *ZFHX3* gene in normal tissue data (https://www.proteinatlas.org/about/download; downloaded 10/10/2023). The HPA data for ZFHX3 are based on a rabbit anti-ZFHX3 polyclonal antibody (Atlas Antibodies #HPA059353, RRID:AB 2,683,987). Protein expression was quantified as "not detected," "low," "medium," or "high." As several 16q-associated ataxias have emerged,[6,9,10] we studied the genetic architecture of this region through functional genomic annotation to gain further biological insights into the region more broadly.[19] Similarly, we used functional genomic annotations to characterize the causative gene as previously described. Specific to the pathogenic mechanisms of these ataxias, we used resources generated through application of Tandem Repeats Finder[20] to the human reference genome (GRCh38) in HipSTR (https://github.com/HipSTR-Tool/HipSTR-references)[21] to create a metric of short tandem repeat (STR) density, size, number of nucleotides within each STR, and location of STRs as annotated by Ensembl v.72[22] across the entire genome. For the gene of interest (*ZFHX3*), we compared the gene-based metric of STR density[19] between genes in which repeat expansions are known to be associated with ataxia, genes associated with ataxia but not through pathogenic expansions, and genes not known to be associated with ataxia, as defined within Genomics England PanelApp[23] (collated gene list available on https://github.com/ZhongboUCL/hereditary-ataxia-functional-genomics). Wilcoxon rank-sum test was used to compare the distributions of these metrics with a two-tailed *P*-value <0.05 deemed to be significant. All analyses were carried out in R (v.4.0.5).

Results and Discussion

Clinical Characterization

The five-generation family were originally from southern Sweden but had migrated to Utah. The family comprised 38 members at the time of the original report after detailed clinical examination by Flanigan and colleagues (Fig. 1A)[6] and later in an extended study of clinical features of SCA.[24] The median age of disease onset was 39.3 years (range: 19–59 years). Patients without symptoms under the age of 40 years had been designated as "unknown" in terms of whether they were affected given the potential for later age-of-onset. More detailed findings are presented in Table 1. Gait ataxia was an early symptom alongside asymptomatic sensory impairment revealed by clinical examination, with vibration sensation loss being the most commonly impaired. All patients had loss of distal lower limb reflexes: 12 of the 13 patients examined on neurophysiology had absent sural sensory nerve action potentials.[6]

Although a fifth of patients had extensor plantar responses, there was no evidence of other pyramidal features, including spasticity or brisk reflexes. Half of patients also had dysarthria.[6] None of the patients had diplopia, but eye signs varied between individuals (Table 1).[24] There was no evidence of autonomic symptoms.

We obtained previously extracted DNA derived from individuals of the family described in the original study and DNA from lymphoblast cell lines derived from four other members of the same family. Unfortunately, DNA extracted as part of the original study of the affected individuals mostly showed variable fragment lengths (range: 1–17 kb) and quantity (0.45–230 ng/μL), with only two samples surviving quality control (17278) and an unrelated spouse (22024) in the Utah pedigree. One sample from the Iowa pedigree was sufficient for long-read sequencing (27376). The analysis therefore used these three samples and an additional three affected probands (III-2 [18956], IV-14 [19138] and IV-20 [23712], Fig. 1A) and an unaffected spouse (IV-12 [17274]) as control, for whom we had lymphoblast cell lines and from which high-molecular-weight DNA suitable for long-read analysis was derived. This amounted to seven samples in total.

Patient III-2 was a nursing home resident at the time of enrolment. She was symptomatic with ataxia. Patient IV-14 reported symptom onset at the age of 59 with progressive loss of balance. He also reported that he had difficulty feeling the pedals of his car when driving with tingling in his hands and legs. Examination 8 years after disease onset revealed a mild dysarthria with mild truncal, gait, and symmetrical limb ataxia. There was also evidence of reduced light touch in the distal upper and lower limbs with reduced vibration sensation in the feet. He was areflexic throughout with upgoing plantar responses (but no other pyramidal signs). Nerve conduction studies showed absent sural and radial sensory responses. Patient IV-11 noticed symptoms at the age of 45 years. Patient IV-20 had a younger age-of-onset. At the age of 25 years, he noticed difficulties with balance, which progressed slowly until he developed problems with fine motor function of his hands in his late 30s due to sensory loss. Examination in his early 40s showed evidence of a mild limb, gait, and truncal ataxia. Vibration sensation and joint position sensation was reduced at the distal interphalangeal joints. Reflexes were absent in the upper limbs and diminished in the lower limbs with flexor plantar reflexes. Both patients had no evidence of oculomotor or other cranial nerve involvement, with normal cognition.

In the second family of Swedish ancestry residing in Iowa (Fig. 1A), Patient II-1 (27376) started to experience a loss of balance and numbness in his legs at the age of 50. Ten years after disease onset, the patient started to need a wheelchair for mobility. There was no
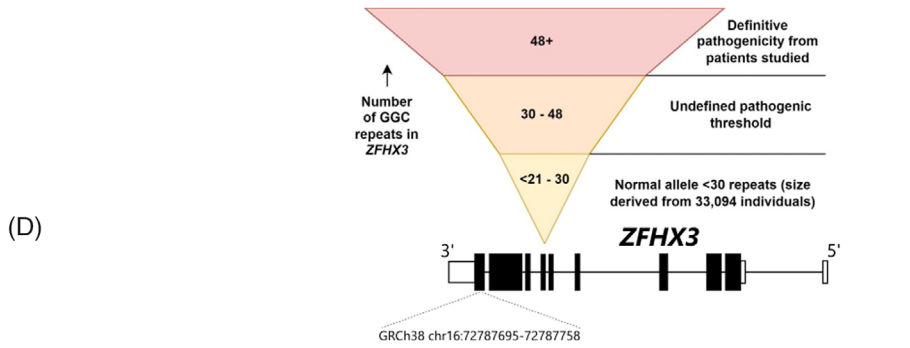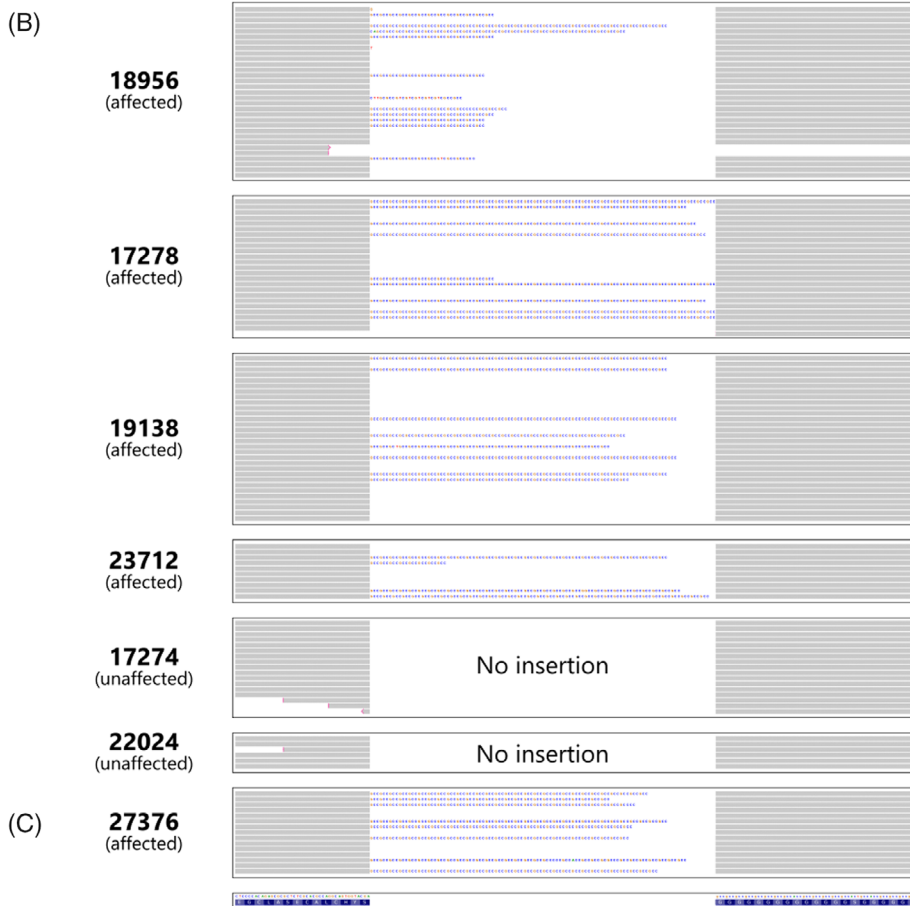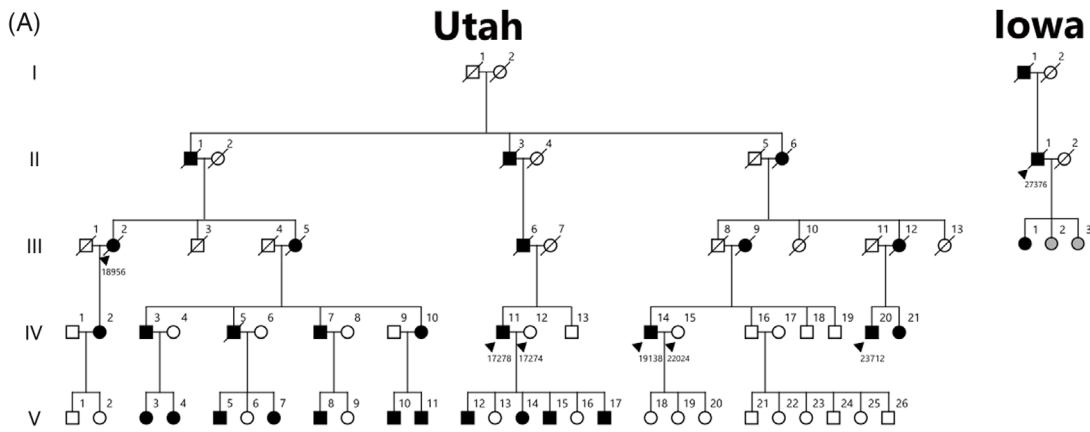
FIG. 1. Legend on next page.

TABLE 1   *Summary of clinical features of SCA4 patients reviewed from the pedigree*

Category	Clinical feature	Proportion of patients affected (%)	n
Cerebellar	Gait ataxia	95	20
	Limb ataxia	95	20
	Dysarthria	50	20
	Hypotonia	0	14
Neuropathy	Reduced sensation	100	20
	Vibration	100	20
	Pinprick	95	20
	Abnormal reflexes	100	20
	Absent ankle jerks	100	20
	Absent knee jerks	85	20
	Complete areflexia	25	20
	Tingling and cramps	50	14
	Absent sural sensory nerve action potential on nerve conduction studies	92	13
Pyramidal	Upgoing plantar reflexes	20	20
	Spasticity	0	14
	Brisk reflexes	0	14
	Distal weakness	20	20
Eye features	Diplopia	0	14
	Irregular saccadic pursuit	57	14
	Nystagmus	7	14
	Reduced eye movement velocity	43	14
	Head-eye lag	29	14
	Vertical hypermetric saccades	57	14
	Horizontal hypermetric saccades	7	14
	Hypometric saccades	0	14
Other	Reduced cough	14	14
	Incontinence	57	14
	Myoclonus	0	14
	Faciolingual fasciculations	14	14

Note: Clinical features were summarized from the 20 patients reviewed in Flanigan et al,[6] and a subset of these patients (n = 14) were studied in further detail in Maschke et al 2005.[24] Clinical features are grouped into broad categories pertaining to symptoms and signs associated with different systems such as cerebellar dysfunction and peripheral neuropathy. n represents the total number of individuals diagnosed with SCA4 clinically for which the clinical feature was assessed.

FIG. 1. (**A**) Family pedigree of the original Utah family described by Flanigan et al[6] and of another family of Swedish origin residing in Iowa. We used adaptive long-read sequencing to investigate structural variation in the candidate region of linkage in 16q22 in those individuals marked with a black arrowhead. (**B**) Detected structural variant of the GGC repeat expansion in *ZFHX3* in SCA4 cases: 17278, 18,956, 23,712, and 19,138, and (**C**) 27,376 in the Iowa family. There is notable absence of the structural variant in 17,274 and 22,024, two unrelated spouses, which is undetected compared to the reference panel. The reference sequence is given in the bottom panel. The visualization is taken from Integrative Genome Viewer using results from variant calling through Sniffles2 output. The number of reads shown here is only of the inserted GGC repeat, in addition to the 21 repeats in the reference genome. (**D**) Schematic showing the GGC repeat size and associated clinical phenotype and location of the repeat expansion in the *ZFHX3* canonical transcript of ENST00000268489.10. In 33,094 individuals assessed in our study, the median GGC repeat size was 21, and below 30 repeats. Repeat sizes equal to and above 48 are pathogenic within our pedigree. A pathogenic threshold remains to be defined between 30 and 48 GGC repeats—more cases are required to decipher whether repeats of this size are pathogenic. [Color figure can be viewed at wileyonlinelibrary.com]

evidence of diplopia, cognitive difficulties, or autonomic symptoms. Examination showed normal cranial nerves. The patient was areflexic throughout with downgoing plantar reflexes. Sensory examination showed reduced joint position sensation to the ankles, reduced light touch to the mid-calves, and reduced vibration sensation in the lower limbs, which were preserved in the upper limbs. He was dysarthric and had limb, truncal, and gait ataxia. The patient reported that his father was similarly affected around a similar age. He also had a daughter who had similar symptoms with onset around 50 years of age. There was little difference between the clinical phenotype of the Utah and Iowa families although genealogy examination did not show a common shared ancestor.

A GGC Repeat Expansion in *ZFHX3* Segregates with SCA4

We employed a targeted long-read DNA sequencing approach to fully explore a 20 Mb region including the region of high linkage in 16q22 in this pedigree. We applied a novel pipeline to detect overlaps and segregation in structural variation within the region of interest. The targeted sequencing resulted in a mean $117,470 \pm 19,042$ reads mapping to the 20 Mb region, with a mean coverage of $99.8\% \pm 0.3\%$ and a mean depth of $21.4x \pm 13.2x$. Using this approach, we identified three heterozygous variants in all three affected individuals, not present in the unaffected sample. The first two variants were intronic deletions: a 176 bp deletion in the *AMFR* gene (chr16: 56397704–56,397,880) and a 75 bp deletion in *LINC00922* (chr16:65537058–65,537,132). Notably, both of these deletions overlapped with "known" structural variations, annotated in the Database of Genomic Variations.[25] Lastly, we identified a $(GGC)_n$ expansion in the last coding exon of *ZFHX3* (chr16:72787695–72787758), which segregated with the disease and was not expanded within the unrelated control. The four affected family members sequenced were heterozygous for 57 repeats (17278; IV-11), 53 repeats (23712; IV-20), 52 repeats (18956; III-2), and 48 repeats (19138; IV-14). For all four patients the other allele had 21 repeat units. The unaffected spouses (17,274; IV-12 and 22,024; IV-15) had 21 repeats on both alleles (Fig. 1B–E). Interestingly, non-expanded alleles had "A" interruptions non-specifically within the 21 repeats, suggesting its potential importance for repeat stability. A summary of the findings is shown in Supplementary Table S1.

GGC Repeat Size in *ZFHX3* in the 100,000 Genomes Project and In-House Database

To establish the distribution of this candidate GGC repeat size within the normal population, we used ExpansionHunter,[17] a bioinformatics tool that estimates repeat sizes of STRs from short-read WGS data and applied it to the 100,000 Genomes Project and also our in-house database, Koios. Repeat size estimates were subjected to quality control by visual inspection as previously described.[26] In WGS data from 7650 probands without neurological presentation (controls), the median GGC repeat size was 21 (range: 18–30). In 14,186 individuals recruited in the 100,000 Genomes Project with a neurological presentation, the median GGC repeat size was 21 (range: 15–31). Similarly, in 803 individuals presenting with ataxia for which a molecular diagnosis has not been found to date, the median repeat size was 21 (range: 17–23). Furthermore, given SCA4 was described in a family of Swedish ancestry, we also compared GGC repeat size at the *ZFHX3* locus across different ancestries, derived from genetically determined ancestry data taken from metadata supplied in the 100,000 Genomes platform. Of the studied individuals, 15,824 were of European ancestry (757 were patients presenting with ataxia), 1728 were of African ancestry (64 had ataxia), 1518 were of American ancestry (44 presented with ataxia), and 131 were of East Asian ancestry (6 presented with ataxia) (Fig. 2A). In summary, we found no differences in the GGC repeat size between unsolved ataxia cases and controls, or between the different populations. In the WGS data of 21,836 individuals, we found no GGC repeat expansion >31 repeats at this locus, supporting the pathogenicity of the repeat sizes of 48–53 found in the Utah pedigree. Likewise, in 11,258 exomes from our internal database Koios, we found no GGC repeat sizes greater than 31 repeats. Furthermore, these results suggest that SCA4 is a rare ataxia and is likely linked to a founder effect within the Swedish population. Thus, this is likely to be of low utility when screening for idiopathic late-onset ataxia cases on an ancestry-agnostic population level. However, these results should also be interpreted with some caution, given that we do not have a positive control for the repeat expansion among the short-read sequencing data. Further orthogonal validation is required.

High Density of Repeat Elements within the Genetic Architecture of 16q22

ZFHX3 protein exhibits widespread expression throughout the body, as evidenced by data from the Human Protein Atlas.[18] In the brain, its expression is notably restricted to only three tissues, each with a distinct cell type specificity (Fig. 2B). In the cerebellum, ZFHX3 demonstrates high expression exclusively within the Purkinje cells. Within the cerebral cortex, its expression is primarily localized to neuronal cells. In the hippocampus, the expression is comparatively lower but also confined to the respective neuronal cells within that region (Fig. 2B). Elucidating the functional
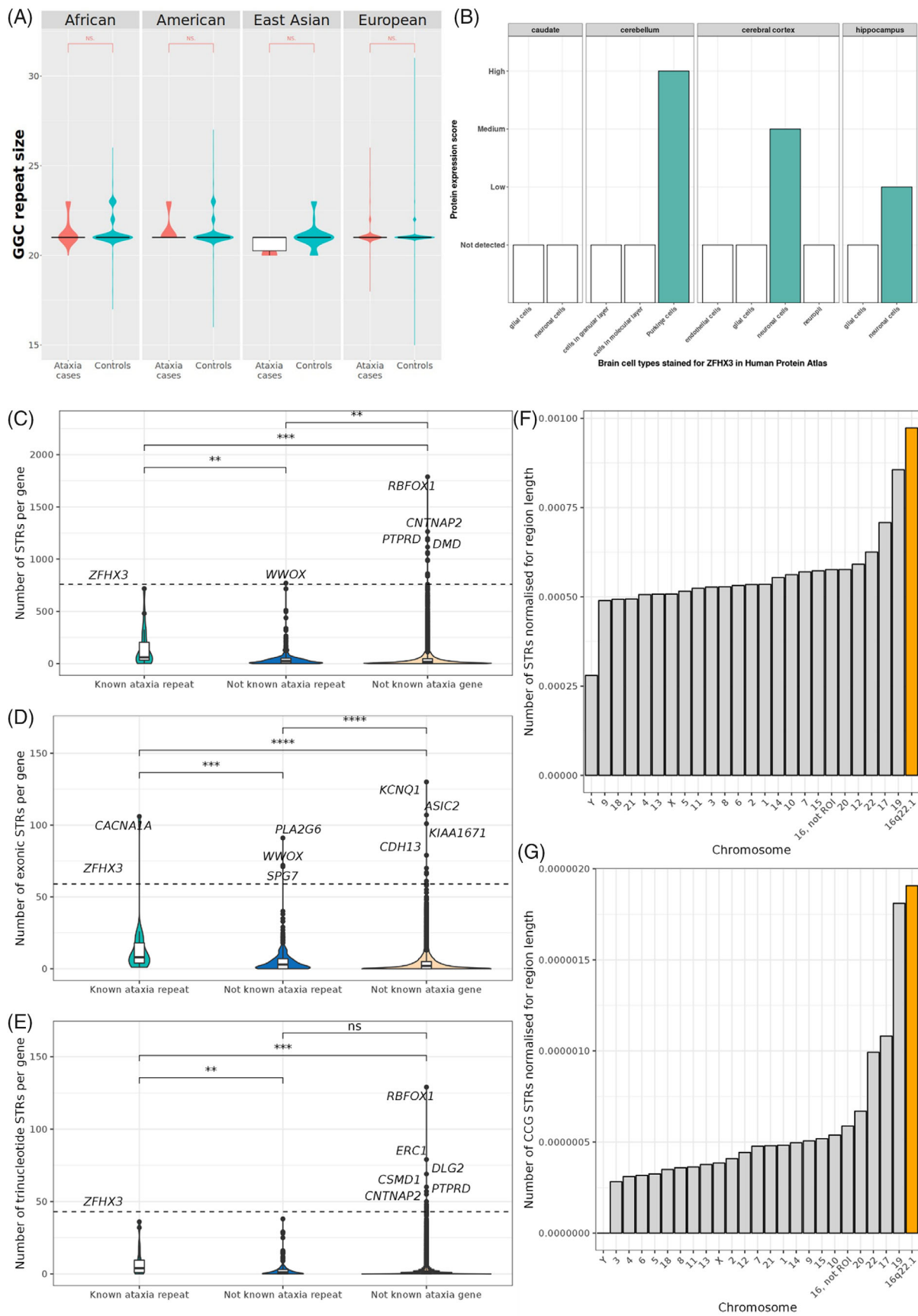
FIG. 2. Legend on next page.

genomic annotation of pathogenic genes has previously been shown to provide insights into the genetic architecture of ataxia and strategies to improve its diagnostic yield.[19] Through this approach, we were able to prioritize potentially pathogenic STRs in an unsolved cohort exemplified by the repeat expansion in *FGF14* associated with SCA27B.[5,27] We predicted that genes harboring a high density of naturally occurring STRs are candidate loci for pathogenic repeat expansions.[19] Here, using a similar approach, we found that remarkably, the *ZFHX3* gene harbors 760 naturally occurring STRs, with an STR genic density that surpasses that of genes in which repeat expansions are known to be associated with disease (Fig. 2C). Moreover, it also has a comparably high density of exonic and trinucleotide STRs compared to other ataxia genes (Fig. 2D,E). This high density of STRs in *ZFHX3* is of functional relevance as genes known to be associated with repeat expansion ataxias have a higher proportion of STRs, trinucleotide STRs, and exonic STRs compared to genes associated with ataxia not known to cause disease through repeat expansion, and protein coding genes not associated with ataxia (Fig. 2C–E). Thus, the gene structure of *ZFHX3* is more akin to a gene in which pathogenic expansion is known to be associated with ataxia. As such, other genes not known to be associated with ataxia but in which there is a high density of STRs (gene symbols annotated in Fig. 2C–E) are potential candidate pathogenic loci to review in unsolved cohorts. Furthermore, we note that the region of interest, 16q22, appears to be enriched for other repeat expansion ataxias.[9,10] Reviewing the genetic architecture of this region, we found that the 16q22.1 region harbors the largest number of naturally occurring STRs and naturally occurring GGC repeats compared to all other chromosomal regions, when normalized for size (Fig. 2F,G). Given the high density of STRs within this region, it is perhaps unsurprising that several repeat expansion disorders have now been characterized as 16q-ataxias.

Conclusions

We describe a heterozygous (GGC)$_n$ repeat expansion in the last coding exon of the *ZFHX3* gene that segregates with late-onset autosomal dominant ataxia associated with sensory neuropathy (SCA4). Although linkage to this region was initially described in 1996,[6] the genetic cause had remained unresolved. Here, the use of targeted long-read DNA sequencing through adaptive sampling proved to be a powerful and efficient approach to successfully solve this diagnostic conundrum.

This approach allowed the accurate detection and sizing of a GC-rich structural variant, a region that we found to be very challenging to amplify through conventional polymerase chain reaction and may therefore underlie the reason for the molecular diagnosis having evaded detection. In addition, this enabled the diagnostic approach to be variant-agnostic; that is, no *a priori* restrictions were placed on the specific repeat motif or type of structural variant during detection. The workflow allowed all structural variants with sequence homology and overlap between the cases to be filtered. In this case, it successfully identified the segregation of the GGC repeat directly as the only plausible pathogenic structural variant candidate. This has obvious advantages to conventional approaches that rely on searching for repeats of a specific motif. In fact, a CAG repeat in *ZFHX3* was already investigated within the German SCA4 family 20 years ago.[8] As the genome harbors approximately 1.7 million polymorphic STR loci,[28] using even more recent bioinformatic tools such as ExpansionHunter[17] requires filtering of variants by type before outlier detection analysis to simplify and manage the number of STR loci reviewed. Thus, adaptive sampling allows for a flexible, computationally directed, variant agnostic approach to target candidate regions of high linkage and represents an exciting and important method in the discovery of unsolved causes of inherited neurological disorders.

FIG. 2. (**A**) Distribution of *ZFHX3* GGC repeat size within 100,000 Genomes Project. The GGC repeat sizes were estimated using ExpansionHunter. The repeat size was compared between controls (defined as participants without neurological presentation) and unsolved ataxia cases. The "NS" above the square brackets indicates no significant differences between the distribution of the repeat size in cases compared to controls as assessed using the Wilcoxon rank-sum test. Ancestries were genetically determined. Of the studied individuals, 15,824 were of European ancestry (757 were patients presenting with ataxia), 1728 were of African ancestry (64 had ataxia), 1518 were of American ancestry (44 presented with ataxia), and 131 were of East Asian ancestry (6 presented with ataxia). (**B**) ZFHX3 protein expression derived from the Human Protein Atlas[18]. (**C**) Distribution of the number of naturally occurring short tandem repeats (STRs) in each gene compared across genes in which pathogenic expansions are known to be associated with ataxia ("known ataxia repeat"), compared with genes associated with ataxia but not through repeat expansion ("not known ataxia repeat"), compared with other protein coding genes not known to be associated with ataxia ("not known ataxia gene"). (**D**) Distribution of the number of naturally occurring exonic short tandem repeats (STRs) in each gene compared across the three gene lists. (**E**) Distribution of the number of naturally occurring trinucleotide short tandem repeats (STRs) in each gene compared across the gene lists. The horizontal dashed line (**C–E**) represents the value for *ZFHX3* for that feature. Annotated genes within the plots show feature value higher than *ZFHX3*. Wilcoxon rank-sum p-values comparing the gene lists are shown as follows: ns: $P > 0.05$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$ (above the square brackets). (**F**) Number of naturally occurring STRs within each chromosomal region, normalized for the size of the region. Chromosome 16 is partitioned into the 16q22.1 region of interest (ROI) and the rest of chromosome 16, without the region of interest ("16, not ROI"). (**G**) Number of naturally occurring GGC STRs within each chromosomal region, normalized for the size of the region. Chromosome 16 is partitioned into the 16q22.1 region of interest (ROI) and the rest of chromosome 16, without the region of interest ("16, not ROI"). [Color figure can be viewed at wileyonlinelibrary.com]

The expansion is associated with a mostly homogeneous clinical syndrome of ataxia and sensory neuropathy (all patients had distal joint position and vibration loss at the time of presentation) across two families of Swedish ancestry. The repeat expansion segregated both with the unrelated spouses and within 33,084 individuals from the 100,000 Genomes Project and from our internal database, Koios. We note that possible anticipation between generations in ages of onset was described when reporting bias was taken into account.[6] Although we recognize that repeat size estimation of a larger number of symptomatic individuals is needed to determine the pathogenic threshold and to assess anticipation, this clinical phenomenon would be in keeping with that of a repeat expansion disorder. Furthermore, it would suggest that different disease severities and presentations could occur within the spectrum of SCA4 with different GGC repeat sizes. With this in mind, we speculate that this cause of ataxia and overlapping sensory neuropathy may be prevalent in Sweden, accounting for another rare cause of inherited late-onset ataxia. Certainly, its rarity within the population suggests that there is a founder haplotype associated with individuals of Swedish ancestry and that screening for this repeat expansion as a common cause of late-onset ataxia in other non-Swedish European populations may have a low yield.

ZFHX3 encodes a large DNA-binding protein that has four homeodomains and 17 zinc finger motifs, with an important function in transcriptional regulation.[29] Thus, it is perhaps unsurprising that *ZFHX3* has been implicated in a range of biological functions and diseases. Common variants in *ZFHX3* are implicated as risk factors for atrial fibrillation[30] and have been reported to modify circadian function through direct interaction with predicted AT motifs in target genes.[31] More recently, loss-of-function variation (no reported variants were in exon 10) has been identified as a novel cause for syndromic intellectual disability.[32] Interestingly, we note that *ZFHX3* shares features with other known spinocerebellar ataxias, being both implicated in transcriptional regulation as in the case of *ATXN2* causing SCA2, and in its Purkinje cell type-specific expression, clearly a cell type of relevance in ataxia. From mice single-cell RNA-sequencing data, *Zfhx3* is an important marker of a ventrolateral spinal cord neuronal subset with long-range afferent projections to the cerebellum.[33] This provides a plausible explanation for both sensory involvement in terms of an early neuronopathy and coexisting cerebellar ataxia in SCA4 that would warrant further investigation.

Importantly, the finding of an exonic GGC repeat expansion strengthens the novel disease entity of polyglycine disorders comprising neuronal intranuclear inclusion disease (NIID) and Fragile X-associated tremor-ataxia syndrome (FXTAS) among others.[34]

Although the other GGC-repeat expansion disorders are not exonic, there is evidence to show translation of the repeat-containing mRNA such as the 5'UTR GGC repeat expansion in *NOTCH2NLC* associated with NIID.[34] Both NIID and FXTAS converge on pathological examination findings of eosinophilic neuronal intranuclear inclusions.[35] With this in mind, it would be of interest to review the pathology of SCA4 for commonality with other GGC repeat expansion disorders.[36] This would also help to inform functional analyses required to determine whether in the case of *ZFHX3*, pathogenicity of the expansion involves a gain or loss-of-function event.

Finally, we note that this study adds to the association of the 16q22 region with repeat expansion ataxias. Through functional genomic annotation, this region was found to harbor a high density of naturally occurring STRs and GGC repeats. Furthermore, we found that *ZFHX3* is a gene with a particularly high density of STRs. These findings provide further insights into the association of causative repeat expansions within the 16q-ataxias and provide a predictive framework for further investigation of structural variants in unsolved cohorts.

These findings were limited by the quality and quantity of DNA from historically old samples. Given the standard required for long-read sequencing, we were unable to gain adequate coverage over the repeat expansion for some of the individuals. Screening in a larger number of individuals would be helpful to characterize whether any anticipation exists.

We have both characterized an exonic GGC repeat in *ZFHX3* associated with SCA4 in the original pedigree in which linkage was characterized over 25 years ago and demonstrated the power of adaptive long-read sequencing to decipher causative structural variation in unsolved cases. ●

Witkowska[1,2]; Suzanne M. Wood[1,2]. [1]Genomics England, London, UK; [2]William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK.

Data Availability Statement

Data available on request due to privacy/ethical restrictions.

References

1. Adam MP, Mirzaa GM, Pagon RA, Wallace SE, Bean LJ, Gripp KW, Amemiya A. GeneReviews® [Internet]; 1993.

2. Klockgether T, Mariotti C, Paulson HL. Spinocerebellar ataxia. Nat Rev Dis Primers 2019;5:24.

3. Sullivan R, Yau WY, O'Connor E, Houlden H. Spinocerebellar ataxia: an update. J Neurol 2019;266:533–544.

4. Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yau WY, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet 2019;51:649–658. https://doi.org/10.1038/s41588-019-0372-4

5. Pellerin D, Danzi MC, Wilke C, Renaud M, Fazal S, Dicaire MJ, et al. Deep Intronic FGF14 GAA repeat expansion in late-onset cerebellar ataxia. N Engl J Med 2023;388:128–141. https://doi.org/10.1056/NEJMoa2207406

6. Flanigan K, Gardner K, Alderson K, et al. Autosomal dominant spinocerebellar ataxia with sensory axonal neuropathy (SCA4): clinical description and genetic localization to chromosome 16q22.1. Am J Hum Genet 1996;59:392–399.

7. Hellenbroich Y, Pawlack H, Rüb U, Schwinger E, Zühlke C. Spinocerebellar ataxia type 4. J Neurol 2005;252:1472–1475. https://doi.org/10.1007/s00415-005-0892-y

8. Hellenbroich Y, Bubel S, Pawlack H, Opitz S, Vieregge P, Schwinger E, Zühlke C. Refinement of the spinocerebellar ataxia type 4 locus in a large German family and exclusion of CAG repeat expansions in this region. J Neurol 2003;250:668–671. https://doi.org/10.1007/s00415-003-1052-x

9. Sato N, Amino T, Kobayashi K, Asakawa S, Ishiguro T, Tsunemi T. Spinocerebellar ataxia type 31 is associated with "inserted" pentanucleotide repeats containing (TGGAA)n. Am J Hum Genet 2009;85:544–557. https://doi.org/10.1016/j.ajhg.2009.09.019

10. Tan D, Wei C, Chen Z, Huang Y, Deng J, Li J, et al. CAG repeat expansion in THAP11 is associated with a novel spinocerebellar ataxia. Mov Disord 2023;38:1282–1293. https://doi.org/10.1002/mds.29412

11. Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. Am J Hum Genet 2021;108:1436–1449. https://doi.org/10.1016/j.ajhg.2021.06.006

12. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res 2021;10:33. https://doi.org/10.12688/f1000research.29032.2

13. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 2018;34:2666–2669. https://doi.org/10.1093/bioinformatics/bty149

14. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191

15. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 2018;15:461–468. https://doi.org/10.1038/s41592-018-0001-7

16. Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, et al. 100,000 genomes pilot on rare-disease diagnosis in health care – preliminary report. N Engl J Med 2021;385:1868–1880. https://doi.org/10.1056/NEJMoa2035790

17. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics 2019;35:4754–4756. https://doi.org/10.1093/bioinformatics/btz431

18. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science 2015;347:1260419. https://doi.org/10.1126/science.1260419

19. Chen Z, Tucci A, Cipriani V, Gustavsson EK, Ibañez K, Reynolds RH, et al. Functional genomics provide key insights to improve the diagnostic yield of hereditary ataxia. Brain 2023;146:2869–2884. https://doi.org/10.1093/brain/awad009

20. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573–580. https://doi.org/10.1093/nar/27.2.573

21. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods 2017;14:590–592. https://doi.org/10.1038/nmeth.4267

22. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res 2021;50:D988–D995. https://doi.org/10.1093/nar/gkab1049

23. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet 2019;51:1560–1565. https://doi.org/10.1038/s41588-019-0528-2

24. Maschke M, Oehlert G, Xie TD, et al. Clinical feature profile of spinocerebellar ataxia type 1-8 predicts genetically defined subtypes. Mov Disord 2005;20:1405–1412. https://doi.org/10.1002/mds.20533

25. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 2014;42:D986–D992. https://doi.org/10.1093/nar/gkt958

26. Ibañez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol 2022;21:234–245. https://doi.org/10.1016/s1474-4422(21)00462-2

27. Chen Z, Tucci A, Ryten M. Deep Intronic FGF14 GAA repeat expansion in late-onset cerebellar ataxia. N Engl J Med 2023;388:e70. https://doi.org/10.1056/NEJMc2301605

28. Jam HZ, Li Y, DeVito R, Mousavi N, Ma N, Lujumba I, et al. A deep population reference panel of tandem repeat variation. bioRxiv 2023. https://doi.org/10.1101/2023.03.09.531600

29. Morinaga T, Yasuda H, Hashimoto T, Higashio K, Tamaoki T. A human alpha-fetoprotein enhancer-binding protein, ATBF1, contains four homeodomains and seventeen zinc fingers. Mol Cell Biol 1991;11:6041–6049.

30. Benjamin EJ, Rice KM, Arking DE, Pfeufer A, van Noord C, Smith AV, et al. Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry. Nat Genet 2009;41:879–881. https://doi.org/10.1038/ng.416.

31. Parsons MJ, Brancaccio M, Sethi S, Maywood ES, Satija R, Edwards JK, et al. The regulatory factor ZFHX3 modifies circadian function in SCN via an AT motif-driven Axis. Cell 2015;162:607–621. https://doi.org/10.1016/j.cell.2015.06.060

32. Del Rocío Pérez Baca M, Jacobs EZ, Vantomme L, Leblanc P, Bogaert E, Dheedene A, et al. A novel neurodevelopmental syndrome caused by loss-of-function of the zinc finger Homeobox 3 (ZFHX3) gene. medRxiv 2023;2023.2005.2022.23289895. https://doi.org/10.1101/2023.05.22.23289895

33. Osseward PJ, Amin ND, Moore JD, Temple BA, Barriga BK, Bachmann LC, et al. Conserved genetic signatures parcellate cardinal spinal neuron classes into local and projection subsets. Science 2021;372:385–393. https://doi.org/10.1126/science.abe0690

34. Boivin M, Deng J, Pfister V, Grandgirard E, Oulad-Abdelghani M, Morlet B, et al. Translation of GGC repeat expansions into a toxic polyglycine protein in NIID defines a novel class of human genetic disorders: the polyG diseases. Neuron 2021;109:1825–1835.e1825. https://doi.org/10.1016/j.neuron.2021.03.038

35. Gelpi E, Botta-Orfila T, Bodi L, Marti S, Kovacs G, Grau-Rivera O, et al. Neuronal intranuclear (hyaline) inclusion disease and fragile X-associated tremor/ataxia syndrome: a morphological and molecular dilemma. Brain 2017;140:e51. https://doi.org/10.1093/brain/awx156

36. Hellenbroich Y, Gierga K, Reusche E, et al. Spinocerebellar ataxia type 4 (SCA4): initial pathoanatomical study reveals widespread cerebellar and brainstem degeneration. J Neural Transm (Vienna) 2006;113:829–843. https://doi.org/10.1007/s00702-005-0362-9

Supporting Data

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

Author Roles

(1) Research project: A. Conception, B. Organization, C. Execution; (2) Manuscript preparation: A. Writing of the first draft, B. Review and critique.

Z.C.: 1A, 1B, 1C, 2A, 2B
E.K.G.: 1A, 1B, 1C, 2A, 2B
H.M.: 1B, 1C, 2B
C.A.: 1B, 1C, 2B
C.C.: 1B, 1C, 2B
C.R.: 1B, 1C, 2B
E.S.: 1B, 1C, 2B
P.A.J.: 1B, 1C, 2B
A.S.: 1B, 1C, 2B
D.P.: 1B, 1C, 2B
K.M.: 1B, 1C, 2B
J.L.: 1B, 1C, 2B
D.G.: 1B, 1C, 2B
H.H.L.: 1B, 1C, 2B
J.H.: 1A, 1B, 1C, 2B
J.P.: 1A, 1B, 1C, 2A, 2B
A.B.S.: 1A, 1B, 1C, 2B
C.B.: 1A, 1B, 1C, 2B
K.D.M.: 1A, 1B, 1C, 2B
A.T.: 1A, 1B, 1C, 2B
Y.H.F.: 1A, 1B, 1C, 2A, 2B
H.H.: 1A, 1B, 1C, 2A, 2B
M.R.: 1A, 1B, 1C, 2A, 2B
L.J.P.: 1A, 1B, 1C, 2A, 2B